The design process of quantitative TaqMan gene expression analysis tools

TaqMan probe-based assays for human and other model organism genes

Abstract

Real-time quantitative PCR (qPCR) has become an established technology for the quantification of gene expression, with the 5´ nuclease assay using Applied Biosystems™ TaqMan® probes [1,2,3] as the gold standard fluorescent reporter method for qPCR. The 5´ nuclease assay using TaqMan probes is sensitive (capable of detecting 10–100 copies of a single transcript in a reaction well) and has a wide dynamic range of detection (transcripts that vary over nine orders of magnitude in mRNA copy number can be detected in a single experimental setup). Another major advantage of this technology is that it is a homogeneous assay in a closed-tube format. The analyte and fluorescent reporter probe are added to a single reaction well and the output signal is read from that well, so no manipulation of the sample is required after it has been added to the PCR mixture. To make this technology accessible to all researchers in a standardized format, we have created an assay design pipeline with the goal of designing TaqMan probe-based assays for all human genes as well as for genes of other model species. This design pipeline integrates both public and proprietary gene sequence information, and uses this information to create the most specific and robust quantitative assays for mRNA transcripts. To date, assays for over 1.3 x 10^6 mRNA transcripts have been designed and released to customers to cover human, mouse, rat, *Arabidopsis*, *Drosophila*, *C. elegans*, and 26 other species.

Introduction

We developed Applied Biosystems™ TaqMan® Gene Expression Assays, a genome-wide collection of quantitative, standardized 5´ nuclease assays for gene expression that enable quantification of gene-encoded transcripts by real-time PCR. The initial goals of the project were to develop at least one assay, based on 5´ nuclease chemistry using TaqMan probes, for all currently known human genes, and to develop assays for many alternatively spliced transcripts of those genes [4,5]. To achieve these initial goals, we developed a probe/primer design pipeline that integrates public, proprietary, and newly discovered human genome information to design gene-specific assays. The resulting assay design pipeline has also been replicated to address 31 other model organism genomes. Additionally, we built a high-throughput oligonucleotide manufacturing facility to produce and inventory the collection of over 2.8 million assays.

The design of this comprehensive set of gene expression assays required that we engineer a completely new and highly sophisticated oligonucleotide probe/primer design pipeline. We accomplished this by developing new, more robust primer design algorithms and an extensive array of bioinformatics tools and processes to automate assay design. The pipeline also integrates design details with the manufacturing process and quality control (QC) of assays (Figure 1).

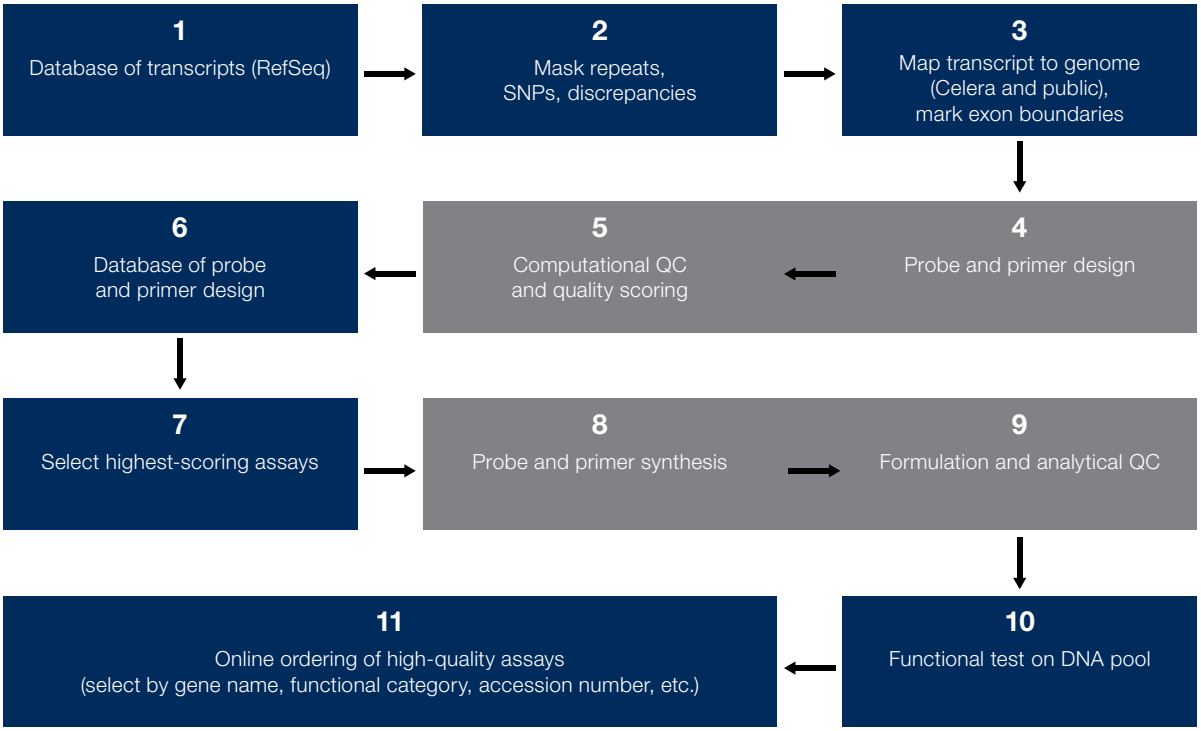

Figure 1. Development pipeline for 5´ nuclease assays to human genes. All our model organism assays are designed with an equivalent development pipeline.

Oligonucleotide probe and primer design for an assay is a critical element in the experimental design process for any real-time PCR experiment. The key feature in assay design is specificity of the assay for the transcript of interest. It is important to ensure that the fluorescence signal being detected is specific to the target transcript, and that there is no contribution from related sequences that might complicate the interpretation of the quantitative results. High target specificity is assured by comparing the sequence of the designed probe and primers to other sequences in the transcriptome from which the gene is transcribed.

It is also important to determine specificity of an assay versus genomic DNA because very often RNA samples can be contaminated with significant amounts of genomic DNA, depending on the RNA purification methods utilized. The genomic DNA contamination problem can be solved with high-quality RNA purification chemistries and rigorous QC of the prepared RNA samples. The transcript specificity problem is more difficult to tackle because of high homology between closely related genes, alternative splicing within a single gene, and the potential presence of transcribed pseudogenes. We have been able to design highly robust assays to deal with many of these specificity issues, and continue to refine our design pipelines to tackle even the most challenging design problems.

Method

Source of transcripts

Transcripts for each of the species we have designed assays for came from the NCBI Reference Sequence Project database (RefSeq [6]; **http:www.ncbi.nlm.nih.gov/RefSeq**), the best known, most highly curated set of nonredundant transcripts in the public domain. We chose this set of transcripts because it is regarded as the most stable set of transcript sequence data available to the public. Each transcript has minimally undergone an automated curation process (Provisional RefSeq Record) and many have gone through a thorough manual curation (Reviewed RefSeq Record) process by NCBI scientists. The majority of transcript sequences in the RefSeq set were derived from cDNA clones, providing good evidence for expression of the transcript, often from multiple sources. Although the assay design process described here can be applied to any set of input transcript sequences, in this paper, we describe the process through which 5´ nuclease assays were designed for the human, mouse, rat, *Arabidopsis*, *Drosophila*, and *C. elegans* RefSeq mRNA sets.

Transcript preprocessing

Each transcript undergoes a preprocessing step that helps to pinpoint the optimal sequence regions within the transcript for designing the oligonucleotide probe and primers for a 5′ nuclease assay. An assay will only be designed in a region of unambiguous sequence that does not contain any known single-nucleotide polymorphisms (SNPs) or repeat sequences. When possible, 5′ nuclease assays for gene expression are designed across exon–exon boundaries, and thus the position of each of the exon boundaries within a multi-exon transcript must be determined prior to the design of each assay.

Transcript preprocessing begins once a batch of transcripts is compiled into a multi-FASTA file. First, repetitive and low-complexity regions in each transcript are masked (i.e., nucleotides are replaced by an "N") using RepeatMasker. Examples of sequences masked at this stage are simple repeats (di- and trinucleotide repeats), *Alu* repeats, SINEs, and LINEs.

Gene structure is annotated by mapping the masked transcripts to the genome assembly with an alignment tool. The positions of each exon–exon boundary are marked for each multi-exon transcript; single-exon transcripts are identified as such. Mapping was performed against the Celera genome assembly, with supplemental mapping information provided by public sequence data, such as reference genomes. If sequence discrepancies are found between the public transcripts and the Celera genome during this step, then the discrepant bases are masked.

In the final preprocessing step, all known SNPs are masked after performing a Basic Local Alignment Tool (BLAST™) [7] analysis against the Celera SNP database and identifying all of the known SNPs within each transcript. We use a proprietary database that integrates SNP information from a variety of sources including dbSNP, the Human Gene Mutation Database (HGMD™) repository, and Celera Discovery Systems as well as information from the Applera Genome Initiative. This database contained over 10 million SNPs at the time the pipeline was developed. Both the SNP-masking and sequence discrepancy–masking steps ensure that no oligonucleotide probe or primer will be designed over ambiguous or known variant nucleotides.

Assay design

Through our extensive experience with probe and primer design of 5′ nuclease assays for quantitative RT-PCR, we have empirically determined the parameters useful for selecting oligonucleotide sequences that are most likely to result in successful, functional assays. We have codified these parameters in a program we call TaqExpress, which is made up of the assay design pipeline and an *in silico* QC pipeline.

In addition, failure analyses have allowed us to recognize oligonucleotide sequences that can be problematic for the generation of robust 5′ nuclease assays. This has allowed us to dynamically engineer the pipeline to eliminate problematic sequences from designed assays.

The gene expression assay design pipeline is an automated process that uses a set of algorithms to design assays to a large number of input transcript sequences. The TaqExpress algorithms are a significant enhancement of the algorithms resident in our Applied Biosystems™ Primer Express™ Oligo Design Software. With the TaqExpress algorithms, we have applied our knowledge and experience in 5′ nuclease assay probe and primer design (including optimal T_m requirements, GC content, buffer and salt conditions, oligonucleotide concentrations, secondary structure, optimal amplicon size, and reduction of primer-dimer formation) to help ensure the design of the most robust assays for the target(s) of interest.

Each of our gene expression assays includes a single TaqMan probe [9] with a minor groove binder (MGB) moiety and two unlabeled oligonucleotide primers. TaqMan MGB probes incorporate both an MGB and a nonfluorescent quencher (NFQ) at the 3′ end of the oligonucleotide. The MGB moiety enhances the T_m by binding in the minor groove of a DNA duplex, enabling the use of shorter TaqMan probes that still meet the higher T_m criteria of a 5′ nuclease probe in conjunction with the lower T_m primers. The nonfluorescent quencher provides greater signal-to-noise ratios, and thus, increased sensitivity of detection of the target transcript over TAMRA™ dye–quenched probes. The use of MGB probes increases the probability of designing an assay in traditionally difficult sequence regions (i.e., AT-rich sequences). Additionally, we have found that the relatively short MGB probes increase the probability that we can design a probe over every exon–exon boundary of a multi-exon gene.

For transcripts from multi-exon genes, an assay target position is selected at each exon–exon boundary. We chose to place the probe, rather than one of the primers, over the exon–exon boundary to ensure that the primers bind in two distinct exons. Placing the probe over the exon–exon boundary ensures that the fluorescent signal is only generated from templates that have correctly spliced exons. All human assays designed over exon–exon boundaries are designated "Hs********_m1", where "Hs" indicates the species, *Homo sapiens*, and the "m" indicates multiple exons.

For single-exon genes, we are limited to placing both the probe and primers within the exon. Assays that have the probe and primers placed within a single exon are therefore designated "Hs********_s1", where the "s" indicates a single exon. This designation was chosen to indicate to users that there is the potential to amplify contaminating genomic DNA in an RNA sample, and that users should implement the appropriate experimental design controls to avoid this problem. The prefix of the Assay ID indicates the species to which the assay was designed (i.e., "Hs" indicates *Homo sapiens*, "Mm" indicates *Mus musculus*, "Rn" indicates *Rattus norvegicus*, etc.).

For multi-exon genes, the pipeline will design up to n – 1 assays (where "n" is the number of exons) across exon–exon boundaries. For transcripts from single-exon genes, multiple assays are also designed by designating target positions that are dispersed across the entire length of the transcript. The design of multiple assays for each transcript provides two advantages: 1) it increases the probability that a successful assay will emerge at the end of the entire design and QC process, and 2) having assays that are designed from the 5′ to the 3′ ends of every transcript provides great flexibility in the choice of a high-quality assay at any position on the transcript.

In silico quality scoring

Downstream from the assay design pipeline is an *in silico* QC analysis pipeline through which every designed probe and primer set is processed. This process penalizes, and thus screens out: 1) assay designs that are not highly specific for the gene of interest, and 2) assay designs that may not accurately report the quantitative expression results for a particular target (i.e., an accurate threshold cycle [C_t] value) in a 5′ nuclease assay [10].

There are three major parts to the *in silico* QC pipeline, and each step generates a penalty score specific to a given assay design. A final penalty score for each assay design comprises the sum of each of the three individual penalty scores. The assay design with the lowest cumulative penalty score for each transcript is the assay that is chosen for manufacturing as a TaqMan Gene Expression Assay.

The three parts of the *in silico* QC process described below in the "BLAST scoring versus sequence databases" section involve:

1. Transcript BLAST scoring

Determining the degree of homology, through BLAST [7], between the assay and other closely related transcripts. A penalty is assigned if an assay detects any closely homologous transcript(s) other than the intended target.

2. Genome BLAST scoring

Determining the degree of homology, through BLAST, between the assay and non-self regions of genomic DNA (i.e., homologous genes and pseudogenes). A penalty is assigned if an assay hits a second or additional physical location on the genome in addition to the location of the target gene.

3. Intron size scoring

Determining the size of the intron across which the probe spans (for assays to multi-exon genes). A penalty is assigned when the assay is designed across an exon–exon boundary that spans a small intron (i.e., <600 bp).

BLAST scoring versus sequence databases

For all BLAST searches, a QC query construct is made by generating an amplicon sequence that includes the intervening probe and each of the two primers; the amplicon is created by padding the specific number of nucleotides between the probe and the primers with “Ns” (Figure 2).

1. Transcript BLAST scoring

The QC query construct for each 5´ nuclease assay is used in a BLAST search against transcript database(s) to ensure that a) each trio of probe and primers in the QC query sequence perfectly matches the target transcript sequence, and b) each assay is specific for the gene of interest and will not amplify transcripts from any other genes.

Primers with homology to other genes (with an intervening homologous probe) can produce an unwanted fluorescent signal, and thus an artificially low C_t value. Primers to homologous genes (without an intervening homologous probe) may amplify homologous transcript(s) besides the target transcript and cause competition for reagents in the PCR mixture, resulting in an artificially high C_t value if the competing homologous transcript is expressed at high levels. These types of side reactions can skew the C_t for the gene of interest and thus produce an erroneous quantitative result for the target transcript. If homology exists, an assay is assigned a penalty score based on the degree of homology to other transcripts. Three sets of numbers are reported in this transcript BLAST step and are described in the bullet points below.

- **BLAST hit to self (Transcript_SelfHSP)**
  The high-scoring pair (HSP) from this BLAST search should produce a match of 100% homology with self. This HSP represents the alignment of the QC query construct to the target transcript in the transcript database. It should show a “0 0 0” (representing zero mismatches in the forward primer sequence, zero mismatches in the probe sequence, and zero mismatches in the reverse primer sequence) result when it is used in a BLAST search against the database from which the target transcript was retrieved (Figure 3).

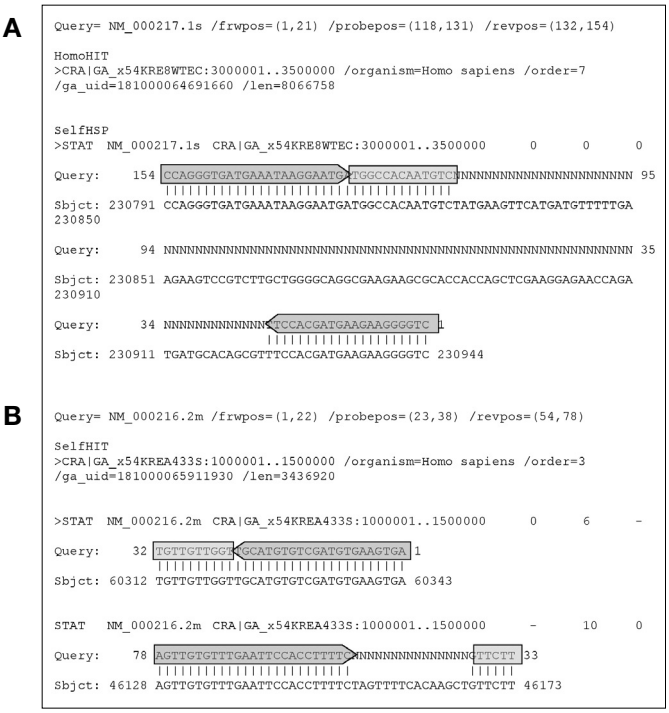

Figure 2. Results of BLAST search versus human genome sequence (Genome_SelfHIT). (A) NM_000217 (*Homo sapiens* potassium voltage-gated channel, shaker-related subfamily, member 1 [episodic ataxia with myokymia] [*KCNA1*]) is a single-exon gene. The probe and primers align perfectly with the genomic DNA sequence. **(B)** NM_000216 (*Homo sapiens* Kallmann syndrome 1 sequence [*KAL1*]) is a 12-exon gene. This assay was designed over the exon 6–exon 7 boundary. The probe sequence is therefore split between the two exons and the intervening intron is ~14 kb long.

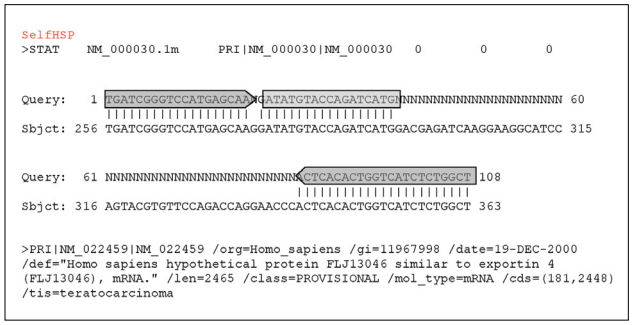

Figure 3. BLAST hit to self transcript (Transcript_SelfHSP). This is an example of a BLAST alignment of two primers and the TaqMan probe sequence against the transcript to which the assay was designed. The primer sequences are shown in the shaded arrows and the probe sequence is shown in the shaded box.

- **Continuous BLAST hits to non-self transcripts** (Transcript_HomoHSP)
  In this set of BLAST results, the top non-self HSPs are reported (i.e., BLAST results to homologous transcripts). The highest penalty is assigned to the HSP that is the closest homolog, but that is not a perfect match to the QC query construct. If two HSPs have the same homology score to the query construct, then the one with the higher homology to the probe region is chosen as the top hit.

- The algorithm will skip all of the homologs that have a "0 0 0" match and will only report the top non-zero HSPs. Therefore, a probe/primer set that can amplify alternative splice variants for the same gene will not be penalized, since these alternatively spliced transcripts may be present as unique transcripts within the database being queried. This step ensures that assays are gene specific, but not necessarily transcript specific. Two or more highly homologous genes may end up with the exact same assay design in regions where the genes have identical sequences. In such a situation, a transcript penalty would not be assigned (because of the "0 0 0" match). This type of situation, where an assay could detect transcripts from more than one gene, is penalized in a downstream part of the in silico QC process, when a BLAST search is done against the genome assembly (see "Genome BLAST scoring" section below). Designing the process in this manner allows us to differentiate between an assay detecting an alternatively spliced variant of the same gene versus an assay that detects a transcript from a different gene locus.

- **Noncontinuous BLAST hits to non-self transcripts** (Transcript_HomoHIT)
  A BLAST query of the probe and primer sets against transcript databases penalizes any alignments to noncontinuous regions of homologous transcripts. This type of BLAST result is called a HIT because the QC query construct hits two different (noncontiguous) parts (HSPs) of a non-self transcript. This type of transcript BLAST result reflects a situation where the amplicon from a homologous transcript would be a different size than the target amplicon. Unlike the previous steps, these results are not from the same HSP, but are from two different HSPs (Figure 4). The higher the homology between the primers and the HIT, the greater the penalty. This penalty is assigned in order to minimize the possibility of nonspecific amplification of transcripts other than the target, and thus, competition for reagents in the PCR mixture that could affect the C_t of the target of interest.

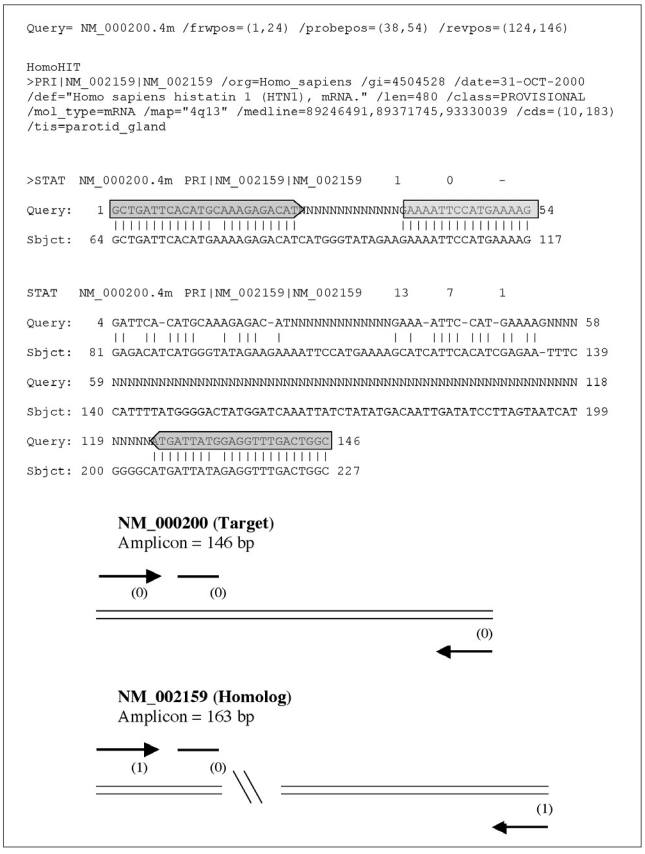

Figure 4. BLAST results showing a significant hit to a non-self transcript (Transcript_HomoHIT). An assay was designed across exon 4–5 of NM_000200. In this example, there is a perfect BLAST alignment to the self transcript (NM_000200, *Homo sapiens* histatin 3 [*HTN3*]; not shown) and a significant alignment to a second (non-self) transcript (NM_002159, *Homo sapiens* histatin 1 [*HTN1*]). The results show that each of the primers has only a single mismatch to the NM_002159 transcript, and that the probe has a perfect match to this non-self transcript. Because of this non-self hit, this particular assay design would be assigned a very high transcript penalty score and would not be manufactured.

2. Genome BLAST scoring

The same QC query construct that is used in a BLAST search against the transcript databases is also used in a BLAST search against the respective species genome assembly, and the output is reported in a very similar manner. This QC step is important because 1) the transcript databases may not be comprehensive, and thus, a homologous transcript that is not yet known could be missed, 2) genomic alignment allows us to distinguish different genes from alternative splice variants of the same gene, 3) it allows us to minimize amplification of artifacts that result from the possible presence of contaminating genomic DNA in a total RNA sample, and 4) it allows us to penalize those probe/primer sets that would amplify pseudogenes in total RNA samples that contain contaminating genomic DNA.

- **BLAST hit to self (Genome_SelfHIT)**
  As with the BLAST search to align the probe and primers to the target sequence in the transcript databases, similar BLAST searches are used to align the probe and primers to the unique gene in the genome to which they were designed. For multi-exon genes, the match must be "0 X 0" for the probe/primer set to avoid a penalty. The two zeros represent no mismatches between the forward and reverse primer sequences and the genome sequence, and the fact that they come from two different HSPs indicates that the primers are on two different exons, separated by an intron. The non-zero value of "X" reflects the fact that the probe is interrupted by an intron, and therefore, does not align itself to a contiguous sequence in genomic DNA. For single-exon genes, the BLAST search alignment should return a value of "0 0 0" because there are no intronic regions to interrupt the probe sequences and lead to mismatches (Figure 2).

- **Continuous BLAST hits to non-self gene(s)** (Genome_HomoHSP)
  The Genome_HomoHSP BLAST results identify genomic regions that have high homology to the probe and primers, and would amplify a PCR product of similar size to the target transcript from contaminating genomic DNA present in an RNA template. This situation would most often occur because of the presence of a pseudogene in genomic DNA. This BLAST result identifies the HSP with the highest homology to the amplicon, with the focus primarily in the two primer regions. If two HSPs have the same degree of homology in the primer sequences, then the HSP with a higher homology to the probe region is chosen as the top hit, and the degree of mismatch in the primers and probe is used to generate the penalty. The higher the degree of homology between the primers and probe and the HSP, the greater the penalty. In a sense, we are over-penalizing assays by assigning this genomic DNA penalty. However, we apply this penalty in order to maximize the ability of an assay to accurately quantitate the target of interest in RNA preparations that may be contaminated with genomic DNA.

- **Noncontinuous BLAST hits to non-self gene(s)** (Genome_HomoHIT)
  This genomic BLAST alignment identifies the genomic sequences that have the highest homology to each of the primers but come from two different HSPs. If the intervening sequence between the two HSPs is short, then the penalty is high. This ensures that we minimize the chance of amplifying a non-target template in an RNA preparation with genomic DNA contamination. If the genomic interval between the two primers is large, the penalty is smaller because it is unlikely the primers would actually produce an amplicon from this type of secondary template.

As previously described, we do not penalize non-self "0 0 0" hits in the transcript BLAST QC step, so we use the Genome_HomoHIT BLAST results to penalize assays that cannot discriminate between homologous genes. If two or more highly homologous genes have identical assays designed (i.e., in a region where the two different genes have identical sequences), then the assays are penalized at this step. If the Genome_HomoHIT results shows "0 X 0" hits in at least one genomic location in addition to self, then the assay is assigned a large penalty because it is assumed that this second hit is to a separate and distinct gene.

3. Intron size scoring

The third part of the *in silico* QC scoring process is the determination of intron size for assays to multi-exon genes that have the probe spanning an exon–exon boundary. Although a penalty for small intron size is integrated into the Genome_HomoHIT rule, a separate rule also penalizes probe/primer sets that span small introns. This reduces the possibility of competition for reagents in RNA samples contaminated with genomic DNA, and also decreases the chance of amplifying incompletely spliced transcripts. The intron penalty is based on the size of the intron: the larger the intron, the smaller the penalty.

Relational database for 5´ nuclease assay designs (TaqDB)

Assays designed for transcripts are all stored in a relational Oracle™ database (TaqDB). The TaqDB database serves as a central repository for all assay designs. It aggregates information about transcripts, assays, global relationships between transcripts and assays, exon–intron structure, *in silico* QC, manufacturing order status, and analytical QC data determined in the manufacturing process. Data from expression studies in select RNA tissue pools are also stored in this database.

Loading information

The TaqExpress assay design and QC pipeline outputs several files, including oligonucleotide sequences, *in silico* QC scores, and BLAST hits, in a flat file format. A database-loading pipeline processes data for each assay into TaqDB.

When a new set of assays is being loaded into TaqDB, the first process is to compare the probe and primer sequences in the incoming design file with those of the assays already in the database. All newly designed assays with less than 100% sequence identity to an existing assay are added as new records into the database. Any incoming assay that has 100% sequence identity with an existing assay will not be added into the database. Instead, a link is created to the existing assay, and any new assay information such as QC score changes or new BLAST results will be updated in the database.

Linking assays to transcripts

A large number of BLAST searches against a variety of databases (i.e., RefSeq, GenBank, Mammalian Gene Collection [MGC], Celera Transcripts [i.e., hCT]) are performed during the assay design process, as previously outlined. Approximately 100 BLAST results are stored for each assay.

The BLAST files that are loaded into TaqDB contain the mismatch information resulting from the comparison of the probe and primers to these various databases. When there is a BLAST file showing a perfect match (0 0 0) to a transcript, then a link is created in the database between the assay and the accession ID of that transcript. When there are additional transcripts that perfectly match the probe and primers, they are also added to the database and “virtually” linked to that particular assay. These links are considered virtual because they are links to transcripts that the assay was not originally designed to detect, but which it will detect.

Alternative splice forms of a particular gene are the most common source of virtual links. Cross-referencing all of the BLAST files with all of the assays in this manner allows us to create many-to-many relationships between assays and transcripts, therefore defining which transcripts an assay may amplify. As a result of this process, an assay can match multiple transcript accession IDs, for example, multiple RefSeq entries. In addition, other BLAST files that contain small mismatches are also loaded into the database and linked to the assay as BLAST QC data.

The assay-to-many-transcripts relationships are displayed in the TaqMan Gene Expression Assays online ordering system so that a researcher has information on all of the transcripts an assay is known to detect, as well as the assay location on each detected transcript.

Remapping

Transcript databases change over time; new transcripts are continually being discovered, and occasionally, entries that were originally thought to be transcripts are found to be faulty and are purged. To keep TaqDB and our collection of ready-to-use assays current, we use BLAST searching to map our assays to the new set of transcripts after a new transcript database is released (i.e., when RefSeq is

updated—approximately every four weeks). This process keeps the information current through the identification of every known transcript that a particular assay can amplify, and it also lets us remove any assay in our collection that no longer maps to the up-to-date transcripts. An additional benefit of the remapping process is that we do not need to design assays for every sequence in every transcript database. Rather, we can often find a link from an existing assay to new sequences, and therefore save time in delivery of assay products to researchers.

Creating products from data

TaqMan Gene Expression Assay manufacturing orders are automatically generated from the assay information in the TaqDB. Given a list of transcript IDs, the database identifies transcripts for which no assay exists. For those transcripts, the database identifies the assay design with the highest score from the TaqExpress *in silico* QC step and automatically sends an order for that assay to our high-throughput global manufacturing facilities. An assay will remain categorized as "ordered" until our manufacturing facilities pass information about assay manufacturing and analytical QC back to the database. If an assay fails in manufacturing (i.e., if one of the oligonucleotides proves difficult to synthesize in high yield), the appropriate failure code is entered into the database, which automatically identifies the next most promising assay design, and sends an order for the new assay to manufacturing. This process helps ensure that an assay is successfully developed for all genes.

Data mining

This database not only has served as a data repository, but also has become a valuable tool for mining information. For example, extracting the oligonucleotide sequences from assays that failed in the manufacturing process (i.e., quantification or analytical QC) has allowed us to compare the problematic sequences and identify commonalities. Certain types of sequences have been discovered that tend to be difficult to manufacture. These types of discoveries have allowed us to make improvements to the assay design process by penalizing oligonucleotides that contain the problematic sequences. In turn, this decreases the failure rate in manufacturing and results in better functional assays.

Results and discussion

Over 2.8 million assays for 32 species have been designed using the assay design process (through July 2019). From these transcripts, >70,000 assays have been manufactured and are held in inventory.

Additionally, the remaining assay designs have been added to our website (**thermofisher.com/taqman**) on a made-to-order basis. These assays cover the different locations across each transcript. There are some RefSeq transcripts across the 32 species for which no order has been sent to manufacturing, and these assays fall into the following categories:

1. No assay designed
2. No designed assay passes the current penalty cutoff
   - Transcript penalty
   - Genome penalty
   - Intron size penalty (multi-exon genes only)

Although many of the assays that do not pass our *in silico* QC standards may be suitable assays under certain circumstances, we have chosen to use especially rigorous standards to avoid manufacturing assays that have the potential to produce difficult-to-interpret quantitative gene expression results. There are a variety of reasons why a designed assay may not be a robust assay for quantitative determination of mRNA transcript levels in a particular RNA sample. Thus, not all of these *in silico* QC steps may be important to all users of an assay. Our aim is to provide the most robust quantitative assays that will fit the requirements of the entire spectrum of sample types and sample preparation methodologies utilized by the broad range of users of a particular assay.

Table 1 provides an example of how our process works, showing all of the original assays designed across the exon–exon boundaries of the human plakophilin 4 (*PKP4*) mRNA (RefSeq ID NM_003628.1). Seventeen assays were designed for this transcript. Of the assays designed, only the top-scoring assay that had no design penalties assigned was sent to manufacturing. However, there are six other candidate assays that met the manufacturing QC cutoff for this particular target that can be chosen if for some reason the top-scoring assay fails somewhere along the downstream manufacturing and functional testing processes.

Of the assay designs that did not pass our *in silico* QC cutoff, one had a mid-level score because it was designed over an intron shorter than 200 bp. The rationale for this penalty score is that if the assay was being used to detect the transcript in a total RNA sample contaminated with genomic DNA, then the contaminating genomic DNA could be co-amplified with the mRNA target, potentially leading to inaccurate quantification of the mRNA template. The likelihood of this occurring is low since the primers are at 900 nM each in the final reaction, and the probe does not detect genomic DNA. Co-amplifying targets that do not bind to the probe will not interfere with quantification when present in small amounts. Such targets are often spiked into a reaction to serve as internal quantification controls [11,12].

Table 1. All designs for a single transcript (NM_003628.1). This table shows the original 17 assays designed for this transcript (*Homo sapiens* plakophilin 4 [*PKP4*]; a 22-exon gene).

RefSeq ID	Assay ID score	Final score	Assay design score	Intron penalty	Intron size	Transcript penalty	Genomic penalty	Status
NM_003628.1	Hs00269305_m1	High	High	0	>10 kb	0	0	Ordered
	Hs00269306_m1	High	High	0	>10 kb	0	0	
	Hs00269307_m1	High	High	0	>10 kb	0	0	
	Hs00269308_m1	Mid	High	High	<200 bp	0	0	
	Hs00269309_m1	High	High	0	>3 kb	0	0	
	Hs00269310_m1	High	High	0	>3 kb	0	0	
	Hs00269311_m1	Low	High	Low	>1 kb	0	High	
	Hs00269312_m1	Low	High	0	>10 kb	0	High	
	Hs00269313_m1	Low	High	0	>3 kb	0	High	
	Hs00269314_m1	Low	High	Low	>1 kb	High	High	
	Hs00269315_m1	Low	High	High	<200 bp	0	High	
	Hs00269316_m1	Low	High	0	>2 kb	0	High	
	Hs00269317_m1	Low	High	0	>3 kb	0	High	
	Hs00269318_m1	Low	High	High	<200 bp	0	High	
	Hs00269319_m1	High	High	0	>2 kb	0	0	
	Hs00269320_m1	High	High	Low	>1 kb	0	0	
	Hs00269321_m1	Low	High	Low	>1 kb	0	High	

Ten of the assays designed to the *PKP4* target received a low final score because the probe/primer sequences for these assays exhibited high homology to at least one other portion of the genome. This penalty signals one of three possible situations: 1) that the domain that these exons encode is conserved and is present in other genes, 2) that there exists at least one pseudogene elsewhere in the genome, or 3) that there is a random sequence at another site in the genome with very high homology to these particular exon sequences.

Regardless of the reason, the potential exists for these low-scoring assays to generate less accurate quantitative results in a total RNA sample contaminated with genomic DNA than in a highly purified RNA sample. This emphasizes the importance of high-quality RNA template preparation upstream of any RT-PCR methodologies.

Although the automated TaqExpress pipeline in its current form can successfully design high-scoring assays for the vast majority of transcripts (i.e., ~85% of the RefSeq transcripts), it is not able to design an assay that passes the *in silico* QC for every transcript. An example of a transcript within the RefSeq set for which assays have been designed but no assay was sent to manufacturing, owing to a high genomic DNA penalty, is the human dihydrofolate reductase (*DHFR*) mRNA.

The *DHFR* gene family consists of one functional gene and at least four intronless (or processed) pseudogenes [13]. In the current pipeline, all the assays designed for the *DHFR* functional transcript were assigned a very heavy genomic DNA penalty because of four high-sequence homology BLAST hits to other regions of the genome (i.e., the pseudogenes). Assay designs that detect both functional genes and nontranscribed pseudogenes are again problematic when an RNA template is contaminated with genomic DNA. In some instances, we have chosen to release these types of assays to manufacturing and flag them as assays that will potentially generate amplicons from contaminating genomic DNA ("Hs********_g1" in an assay name indicates the potential for an assay designed over an exon–exon boundary to give a positive signal with genomic DNA). Users of these assays should ensure the purity of their total RNA sample by eliminating genomic DNA contamination (i.e., treating samples with DNase and performing an RT^+/RT^- experiment).

The most difficult situation in which to accurately measure expression of the transcript from the known functional gene is when a pseudogene is actually transcribed. *CYP2D6* is a perfect example of this situation [14].

The *CYP2D6* gene has two known transcribed pseudogenes (*CYP2D7P* and *CYP2D8P*), and these must be considered competing transcripts. The only way to design an assay specific for the transcribed functional *CYP2D6* gene is to design assays in a more targeted fashion through, for example, a pipeline that designs assays based on sequence differences in a multiple sequence alignment. In the case of *CYP2D6*, the multiple alignment must include the sequences of the transcribed pseudogenes. The sequences of transcribed pseudogenes are not always present in transcript databases, and so must be fetched in a case-by-case analysis. This pipeline is currently in use for highly homologous transcripts.

Through the use of this automated assay design and QC pipeline, we have been able to develop simple-to-use 5′ nuclease-based assays for the majority of known genes across a number of species. The examples provided herein highlight the strengths and limitations of this process. The ability to mine the TaqExpress pipeline data for information about why assays do not pass QC and manufacturing has enabled us to continually make improvements to our processes.

For more information on TaqMan Gene Expression Assays, please go to **thermofisher.com/taqmangeneexpression**.

applied
biosystems

Authors

Katherine Lazaruk, Yu Wang, Jennifer Zhong, Sergei Maltchenko, Steven Rabkin, Kathryn Hunkapiller, Manohar Furtado, Olga Petrauskene, Karl Guegler, Dennis Gilbert, and Eugene Spier.

Acknowledgments

We are indebted to Ryan Koehler and Alex Dubman for their contributions to the development of the bioinformatics pipeline and TaqDB database. The authors thank Criss Walworth, Mark Wechser, Kathleen Shelton, and Manohar Furtado for helpful discussions and Mignon Fogarty for assistance with the manuscript.

References

1. Dumur CI, Dechsukham C, Wilkinson DS et al. (2002) Analytical validation of a real-time reverse transcription-polymerase chain reaction quantitation of different transcripts of the Wilms' tumor suppressor gene (WT1). *Anal Biochem* 309(1):127–136.
2. Heid CA, Stevens J, Livak KJ et al. (1996) Real-time quantitative PCR. *Genome Res* 6(10):986–994.
3. Winer J, Jung CK, Shackel I et al. (1999) Development and validation of real-time quantitative reverse transcriptase-polymerase chain reaction for monitoring gene expression in cardiac myocytes *in vitro*. *Anal Biochem* 270(1):41–49.
4. Holland PM, Abramson RD, Watson R et al. (1991) Detection of specific polymerase chain reaction product by utilizing the 5′—>3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc Natl Acad Sci USA* 88(16):7276–7280.
5. Lee LG, Connell CR, Bloch W (1993) Allelic discrimination by nick-translation PCR with fluorogenic probes. *Nucleic Acids Res* 21(16):3761–3766.
6. Pruitt KD, Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29(1):137–140.
7. Altschul SF, Gish W, Miller W et al. (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
8. De La Vega FM, Dailey D, Ziegle J et al. (2002) New generation pharmacogenomic tools: a SNP linkage disequilibrium map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies. *Biotechniques* Suppl:48–50, 52, 54.
9. Livak KJ, Flood SJ, Marmaro J et al. (1995) Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl* 4(6):357–362.
10. Applied Biosystems (2002) Essentials of Real-Time PCR. Foster City, California: Applied Biosystems (p/n 4371089).
11. Furtado MR, Callaway DS, Phair JP et al. (1999) Persistence of HIV-1 transcription in peripheral-blood mononuclear cells in patients receiving potent antiretroviral therapy. *New Engl J Med* 340(21):1614–1622.
12. Mulder J, McKinney N, Christopherson C et al. (1994) Rapid and simple PCR assay for quantitation of human immunodeficiency virus type 1 RNA in plasma: application to acute retroviral infection. *J Clin Microbiol* 32(2):292–300.
13. Anagnou NP, Antonarakis SE, O'Brien SJ et al. (1988) Chromosomal localization and racial distribution of the polymorphic human dihydrofolate reductase pseudogene (DHFRP1). *Am J Hum Genet* 42(2):345–352.
14. Endrizzi K, Fischer J, Klein K et al. (2002) Discriminative quantification of cytochrome P4502D6 and 2D7/8 pseudogene expression by TaqMan real-time reverse transcriptase polymerase chain reaction. *Anal Biochem* 300(2):121–131.

Find out more at
thermofisher.com/taqmangeneexpression

ThermoFisher SCIENTIFIC